DOCUMENT RESUME

ED 453 259 TM 032 796

AUTHOR Rotou, Ourania; Elmore, Patricia B.; Headrick, Todd C.
TITLE Number Correct Scoring: Comparison between Classical True Score Theory and Multidimensional Item Response Theory.
PUB DATE 2001-04-12
NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Item Response Theory; *Scoring; Standardized Tests; Test Items; *True Scores
IDENTIFIERS *Classical Test Theory; *Number Right Scoring; Weighting (Statistical)

ABSTRACT

This study investigated the number-correct scoring method based on different theories (classical true-score theory and multidimensional item response theory) when a standardized test requires more than one ability for an examinee to get a correct response. The number-correct scoring procedure that is widely used is the one that is defined in classical true-score theory (CTT). In CTT, a test score is equal to the number of items an examinee answered, so that all items are weighted "one." It is also possible to use a form of number-correct scoring in which the weights of items are different. In this study, the accuracy of estimated number-correct scores relative to true number-correct scores under CTT, multidimensional item response theory (MIRT) and both MIRT and CTT were studied using simulated data for a standardized test in which true scores and estimated scores were known. A method in which item weights were based on MIRT and test scores based on CTT (MIX method) was found to be the most accurate method used to estimate the true score on an examinee. This MIX method was also significantly different from the other three scoring methods using the bootstrap analysis. An appendix contains definitions of the notations representing the various parameters and approaches. (Contains 20 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 453 259

Number Correct Scoring: Comparison between Classical True Score Theory and Multidimensional Item Response Theory

Ourania Rotou, Patricia B. Elmore & Todd C. Headrick
Southern Illinois University at Carbondale
Department of Educational Psychology and Special Education
Carbondale, IL 62901-4618

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
O. Rotou
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

TM032796

Paper presented at the annual meeting of the American Educational Research Association, April, 2001, Seattle.

BEST COPY AVAILABLE

Number Correct Scoring: Comparison between Classical True Score Theory and Multidimensional Item Response Theory

1. Introduction

From the age that children learn how to read and write, paper and pencil tests play an important role in their lives. Every year, more than 100 million standardized tests are administered in America's public schools (Weaver, 2000). Standardized tests include intelligence tests, achievement tests, career interest inventories, and psychological inventories among others. All children entering kindergarten participate in standardized "readiness" tests that help determine whether a child is ready for the kindergarten program (Pierce, 2000).

Test scores are very important to students, parents, teachers, administrators and professionals. Test scores provide valuable information to students in terms of continuing their education beyond high school. Students can use test scores to select post-secondary institutions that warrant their consideration and perhaps their eventual application. Admission professionals can use test scores to compare students from different states, schools and academic backgrounds. Professionals, in their attempt to better estimate an examinee's ability on a specific trait, develop different scoring methods to derive test scores. A test score is a composite of item scores. Item scores or item weights are the points that an individual would be awarded for a correct response to an item (Frary, 1989).

2. Purpose of the study

A direct result of an examinee's performance on a standardized test is to rank order the individual (according to level of ability) relative to others who took the same test or a parallel test. Test scores are used as estimates of individuals' levels of ability.

The way that test scores are obtained plays a significant role in the outcome of the ranking of individuals. For example, an observed test score of an examinee based on CTT might be different than the observed test score of an examinee based on MIRT and may result in different ranking of individuals and in different decisions.

The goal of this study was to investigate the number correct scoring method based on different theories (classical true-score theory and multidimensional item response theory) when a standardized test requires more than one ability for an examinee to get a correct response. The number correct scoring procedure that is widely used is the one that is defined in Classical True-score Theory (CTT). In CTT a test score is equal to the number of items an examinee answered correctly (NC_{ctt}) (Stocking, 1996). Thus, all items are weighted one.

A second method that utilizes the number correct scoring method is the case in which the weights of the items are different. Theoretically, items within a test are different and provide different information and therefore items should be weighted differently. Particularly, Birnbaum proposed that the weight of an item be equivalent to the item's point biserial value (Lord and Novick, 1968) as defined in CTT. The number correct test score is the sum of the weights of the items an examinee answered correctly (NC_{wctt}).

In theory, the assumption of unidimensionality in Item Response Theory (IRT) is met when there is only one dominant trait (Hampleton, R., Swaminathan, H. and Rogers, J., 1991). In practice, there is more than one trait that may influence an examinee's response such as solving a mathematics word problem that requires two dominant traits, mathematics ability and verbal ability, for an examinee to answer the problem correctly.

A third method that has been utilized is the number correct procedure under the theory of multidimensional item response theory in which a test score is computed as the sum of the probabilities of success (NC_{mirt}). The parameters under investigation in the MIRT model proposed by Reckase (1986) are the multidimensional item difficulty parameter, D_i; multidimensional item discrimination parameter, $MDISC_i$; the angular direction of an item, α, and a vector of abilities $(\theta_{1i}, \theta_{2i})$ of an individual who responds to the item. In a two-dimensional plane an item can be represented as a vector and the item difficulty is the distance from the origin to the point in the space where the item has the steepest discrimination. The item discrimination is the length of the vector and it can be computed using the vector of item discrimination (a_{1i}, a_{2i}) where a_{1i} represents the discrimination of an item i in dimension one and a_{2i} represents the discrimination of an item in dimension two. The angular direction of an item, α, provides the number of degrees an item is from dimension one (Ackerman, 1994).

A problem arises when a standardized test is designed to measure one ability but a second ability is required for an examinee to obtain the correct response. In other words, there is only one dominant dimension but a second dimension is present and it may affect the response of an examinee. For example, the mathematics portion of the SAT test measures one dominant ability, mathematics skill, but a second ability, verbal skill, is required for an examinee to answer the item correctly. One way to resolve this problem is by weighting items based on the skill of interest while controlling for the remaining skill composites. This paper proposes a method that provides item weights based on the dimension of interest while controlling for the second irrelevant dimension. The weights of the items are derived from the formula proposed in this paper that utilizes item

parameters (discrimination index and degree of angular direction) defined in terms of MIRT. Number correct test scores is the sum of the weights of the items that an examinee answered correctly. In other words, item weights are based on MIRT and test scores are based on CTT (NC_{mix}).

This paper investigates the accuracy of the estimated number correct scores, $\hat{NC}$, relative to the true number correct scores under the theories of CTT, MIRT and both CTT and MIRT for a test that measures one ability while a second ability is required for a correct response.

3. Methodology

This study utilized simulated data in which true scores and estimated scores were known to allow for comparisons between true values and estimates (Way, Ansley, Forsyth, 1988).

Selection of Parameters

Most standardized tests are developed to measure one dimension. At the same time, it is realistic that a second dimension is present. Since test items are written primarily to assess dimension 1, it seems reasonable that these items will discriminate more in dimension 1 than in dimension 2 and at the same time the location of the items should be closer to dimension 1 than dimension 2 (Ansley, Forsyth, 1985). With this rationale in mind, parameters for this study were selected based on the following criteria:

(1) Simulated data, abilities of 1000 examinees, (θ_1, θ_2) and item parameters for three test lengths (15-item, 30-item and 50-item), were generated to be a realistic representation of actual test data;

(2) Test items measured one main trait but responses to the items required some skills of a second trait. In other words, the test was designed to measure a primary trait (dimension of interest) but at the same time a second trait was necessary for an examinee to provide a correct successful response to an item. In this paper dimension 1 is represented by the x-axis and dimension 2 is represented by the y-axis.

(3) Data were generated to fit the multidimensional two-parameter logistic model that was developed by Reckase (1986):

$$P(x_{ij}=1|a_i,d_i,\theta)=\frac{1}{1+e^{d_i-\sum_{i=1}^{k}a_{ik}\theta_{jk}}} \qquad (1)$$

where x_{ij} is the response (1 or 0) to item i by person j, θ_{jk} is the ability parameter for person j in dimension k, a_{ik} is the discrimination parameter for item i in dimension k, and d_i is a scalar variable that is linearly related with the difficulty parameter for item i.

Estimation of Parameters

The TESTFACT program (Wilson, Wood and Gibbons, 1998) was used to estimate item parameters and examinee abilities.

Procedure

The procedure described in this section will be repeated 100 times for each of the three tests. For each repetition the seed numbers will be changed.

1. Item and examinee parameters were generated to represent realistic data.

   Alpha, α_i, was generated from a uniform distribution in the interval $[0, \pi/4]$. The vector of the discrimination values of the items in dimension 1, $a_{I.}$ was originally

generated from a uniform distribution [0,1]. It was then rescaled such that a_1 had a mean of 1.23 and a standard deviation (SD) of .34 (Way, Ansley, & Forsyth, 1988). The discrimination of the items in dimension 2, a_2, were computed using (Reckase, 1986)

$$\alpha_i = \tan^{-1}(\frac{a_{2i}}{a_{1i}}) \qquad (2)$$

Where a_{1i} is the item discrimination for item i in dimension 1, and a_{2i} is the item discrimination for item i in dimension 2. The multidimensional discrimination of an item, $MDISC_i$, was computed by (Reckase, 1986)

$$\text{MDISC}_i = \sqrt{\theta_{1i}^2 + \theta_{2i}^2} \qquad (3).$$

Finally, test items were developed to measure a specific set of abilities (θ_1, θ_2). Item difficulty in dimension 1 was set at .7 and in dimension 2 was set at .3, Therefore, the multidimensional difficulty for all items was 0.7615; The scalar variable, d_i that is linearly related to the item difficulty was calculated by the product of the value of the item multidimensional discrimination and the value of the item multidimensional difficulty. The examinees' parameters, θ_1 and θ_2, were generated from a standard normal distribution.

2. Using Reckase's formula (equation 1) for the M2PL model, the probability of a correct response on an item for each examinee was computed, p_{ji} (the probability of examinee j to get item i correct). These probabilities were presented in a matrix NxK_{mirt}, where there were N rows (number of examinees) and K columns (number of items). The entries of the NxK_{mirt}-matrix were the probabilities of an examinee having a correct response p_{ji} (the probability of examinee j to get item i correct). True

Number Correct scores for the examinees under MIRT were calculated by summing the rows of the NxK_{mirt}-matrix. A row represents the responses of a particular examinee. So, at this step the NC_{mirt} score for each examinee was computed.

3. A random number matrix, u_i, from a uniform distribution in the range [0,1] will be used as a comparison matrix. An NxK_{ctt} matrix was formed with indices x_{ji}, by using the following rule:

$$x_{ji} = 1 \text{ if } p_{ji} \geq u_i$$

$$Or$$

$$x_{ji} = 0 \text{ if } p_{ji} < u_i$$

The true number correct score for each examinee, NC_{ctt}, under the CTT was computed by adding the indices of the rows of the NxK_{ctt} matrix.

4. Using the NC_{ctt}, and the NxK_{ctt} matrix the item discrimination values (r_i) can be calculated by (Allen and Yen, 1979, pp. 122): $r_{ix} = \frac{\overline{X}_i - \overline{X}}{S_x}\sqrt{\frac{P_i}{1-P_i}}$, where $\overline{X}_i$ is the mean of the X scores among examinees passing item i, $\overline{X}$ and S_x are the mean and standard deviation of the X score among all examinees, and P_i is the proportion of examinees who answered the item correctly. The NxK_{wctt} matrix can be formed by multiplying the columns of the NxK_{ctt} matrix by r_i. The sum of the rows of the NxK_{wctt} matrix represents the true number correct scores under the weighted CTT (NC_{wctt}).

5. An $N\times K_{mix}$ matrix was formed by multiplying the columns of the NxK_{ctt} by w_i, the weight of item i. Item weight is a function of both the item multidimensional discrimination and the location of the item in the space of the two abilities, θ_1 and θ_2.

The greater the value of the multidimensional discrimination of an item, the greater the value of its weight as long as all other variables are held equal. The location of the item will affect the item's weight as follows: The closer an item is to the dimension of interest (x-axis), the greater the weight assigned given that all other variables are held equal. Alpha, α, called the angular direction of the item, measures how close an item is to the dimension of interest and is computed using equation 2 and the item multidimensional discrimination is computed using equation 3. The new formula that assigns weights to the items is:

$$W_i = \frac{(\Pi/2 - \alpha_i)\text{MDISC}_i}{\sum W_i}.K, \qquad (4)$$

where K is the number of items in the test.

The true number correct score for each examinee, NC_{mix}, under the CTT will be computed by adding the indices of the rows of the NxK_{mix} matrix.

6. The NxK_{cm}-matrix will be used as the input file in the TESTFACT program. The TESTFACT program will provide estimates of the examinee parameters and estimates of the item parameters (e.g. $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{a}_1$, $\hat{a}_2$, $\hat{b}$ and $\hat{d}$).
7. The study repeated the same procedure from step 2 to step 5 but instead of the parameters generated in step one, the estimated parameters from step 6 were used. Using the estimated parameters (from step 6), the following information were obtained:
   (a) Step 2 provided the $N\hat{x}K_{mirt}$-matrix. The entities of this matrix were probabilities of success (based on the estimated parameters) of examinees on test items and the

sum of the rows of the $N\hat{x}K_{mirt}$ matrix provided the estimated number correct scores ($\hat{N}C_{mirt}$) of examinees under the MIRT.

(b) Step 3 was used to create the $N\hat{x}K_{ctt}$ matrix and the sum of its rows provided the estimated number correct scores for the traditional method under CTT ($\hat{N}C_{ctt}$).

(c) Step 4 provided the $N\hat{x}K_{wctt}$ -matrix and by summing the rows of this matrix the estimated number correct score under weighted CTT were obtained ($\hat{N}C_{wctt}$).

(d) Step 5 provided the $N\hat{x}K_{mix}$ matrix and the estimated number correct scores using the formula that assigns weights to items under CTT ($\hat{N}C_{mix}$).

Analysis

Each examinee had eight scores: (1) True number correct score for the traditional method under CTT (NC_{ctt}), (2) estimated number correct score for the traditional method under the CTT ($\hat{N}C_{ctt}$), (3) true number correct for weighted items based on their point biserial correlation value (NC_{wctt}), (4) estimated number correct score based on the weighted items under CTT ($\hat{N}C_{wctt}$),(5) true number correct score under MIRT (NC_{mirt}), (6) estimated number correct score under MIRT ($\hat{N}C_{mirt}$), (7) true number correct score for the method that used the new formula (equation 4) to assign weights to items under CTT (NC_{mix}) and (8) estimated number correct score for the method that utilized the new formula to assign weights to items under the CTT ($\hat{N}C_{mix}$). Comparisons of the form

$$AD = | \hat{N}C - NC |$$

were made using absolute differences (Ansley and Forsyth, 1985).

This study investigated the following absolute differences: $NC_{ctt} - \hat{N}C_{ctt}$; $NC_{wctt} - \hat{N}C_{wctt}$; $NC_{mirt} - \hat{N}C_{mirt}$ and $NC_{mix} - \hat{N}C_{mix}$. Because these absolute deviates are based on different metrics, the coefficient of variation (CV) was used as the standardized measure of comparison (Howell, 1997). The CV is the standard deviation of the absolute deviates divided by the average of the absolute deviates (AAD). There are 100 Coefficients of Variation for each scoring method and for each test length.

Pearson product-moment correlation coefficients were used to better understand the relationships between true *NC* scores and their estimated number correct scores along with other number correct scores . Graphs and tables were presented to show the results of the correlations between variables for the three test lengths (15, 30 and 50-item tests). Tables 1, 2 and 3 present the mean, mode(s) median standard deviation and the range of correlations for the 100 repetitions for the three test lengths for the relationships between (a) true scores, (b) estimated scores and (c) true scores with estimated scores, respectively. Table 4 presents a summary table of the number of samples (from total of 100 samples) that have the smallest coefficient of variation (CV). The smaller the CV, the less variability between estimated scores and the true scores and the more accurate the estimated score. Finally, 95% confidence intervals were constructed using bootstrap techniques (Efron & Tibshirani, 1998) to test for significant differences between the means of the coefficient of variation between the four scoring methods. All bootstrap confidence intervals were based on B=1000 replications (Efron & Tibshirani, 1998, p. 162).

4. Results

Tables 1, 2 and 3 present the descriptive statistics of the correlations for relationships between (a) true number correct scores, (b) estimated number correct scores, and (c) true number correct scores and estimated number correct scores. Tables 1, 2 and 3 show mean, median, mode in parentheses, standard deviation and range for the correlations. The number adjacent to the mode is the frequency of samples at the mode. In general, as the number of items increased the mean of the correlation increased. Also, as the number of items increased the standard deviation decreased and the range decreased.

Table 1
Summary Table of the Correlations between
True Number Correct Scores

Relationship		15-item	30-item	50-item
NC_{ctt} and NC_{wctt}	Mean	.9899	.99	.99
	Median	.99	.99	.99
	Mode	(.99)-99	(.99)-100	(.99)-100
	Stand. Dev.	.001	.0000	.0000
	Range	.98-.99	.99	.99
NC_{ctt} and NC_{mirt}	Mean	.9642	.9767	.9847
	Median	.97	.98	.99
	Mode	(.97)-42	(.98)-60	(.99)-55
	Stand. Dev.	.0153	.0084	.0065
	Range	.91-.98	.95-.99	.96-.99
NC_{ctt} and NC_{mix}	Mean	.9897	.99	.99
	Median	.99	.99	.99
	Mode	(.99)-97	(.99)-100	(.99)-100
	Stand. Dev.	.0017	.0000	.0000
	Range	.98-.99	.99	.99
NC_{wctt} and NC_{mirt}	Mean	.9583	.9743	.9812
	Median	.97	.98	.98
	Mode	(.97)-41	(.98)-57	(.98)-55
	Stand. Dev.	.0203	.0103	.0076
	Range	.86-.98	.94-.99	.96-.99
NC_{wctt} and NC_{mix}	Mean	.9883	.99	.99
	Median	.99	.99	.99
	Mode	(.99)-86	(.99)-100	(.99)-100
	Stand. Dev.	.0045	.0000	.0000
	Range	.97-.99	.99	.99
NC_{mirt} and NC_{mix}	Mean	.9598	.9743	.9834
	Median	.97	.98	.98
	Mode	(.97)- 40	(.98)-53	(.98)-48
	Stand. Dev.	.0190	.0099	.0068
	Range	.86-.98	.94-.99	.96-.99

Table 2
Summary Table of Correlations between Estimated Number Correct Scores

Relationship		15-item	30-item	50-item
$\hat{N}C_{ctt}$ and $\hat{N}C_{wctt}$	Mean	.9712	.9709	.9709
	Median	.97	.97	.97
	Mode	(.97)-86	(.97)-91	(.97)-91
	Stand. Dev.	.0035	.0028	.0028
	Range	.96-.98	.97-.98	.97-.98
$\hat{N}C_{ctt}$ and $\hat{N}C_{mirt}$	Mean	.9178	.9305	.9474
	Median	.93	.94	.95
	Mode	(.97)-15	(.97)-15	(.96)-18
	Stand. Dev.	.0613	.0364	.0263
	Range	.66-.99	.82-.99	.86-.99
$\hat{N}C_{ctt}$ and $\hat{N}C_{mix}$	Mean	.9720	.9715	.9705
	Median	.97	.97	.97
	Mode	(.97)-75	(.97)-85	(.97)-92
	Stand. Dev.	.0060	.0035	.0025
	Range	.96-.99	.97-.98	.97-.98
$\hat{N}C_{wctt}$ and $\hat{N}C_{mirt}$	Mean	.9138	.9294	.9396
	Median	.94	.94	.94
	Mode	(.96)-12	(.97)-16	(.96)-18
	Stand. Dev.	.0727	.0511	.0286
	Range	.60-.99	.60-.99	.84-.99
$\hat{N}C_{wctt}$ and $\hat{N}C_{mix}$	Mean	.9758	.9792	.9813
	Median	.98	.98	.98
	Mode	(.98)-70	(.98)-77	(.98)-72
	Stand. Dev.	.0154	.0076	.0059
	Range	.90-.99	.95-.99	.96-.99
$\hat{N}C_{mirt}$ and $\hat{N}C_{mix}$	Mean	.9079	.9173	.9207
	Median	.925	.93	.92
	Mode	(.96)-11	(.97,.93)-13	(.95)-16
	Stand. Dev.	.0636	.0482	.0355
	Range	.70-.99	.78-.99	.83-.98

Table 3
Summary Table of the Correlations Between
True Number Correct Scores and Estimated Number Correct Scores

Relationship		15-item	30-item	50-item
NC_{ctt} and $\hat{N}C_{ctt}$	Mean	.9106	.9184	.9309
	Median	.93	.93	.94
	Mode	(.94)- 17	(.95)-16	(.94)-25
	Stand. Dev.	.0529	.0361	.0234
	Range	.73-.97	.81-.97	.87-.97
NC_{ctt} and $\hat{N}C_{wctt}$	Mean	.9264	.9275	.9373
	Median	.94	.94	.94
	Mode	(.96)- 17	(.96)-22	(.95)-25
	Stand. Dev.	.0730	.0349	.0248
	Range	.59-.98	.82-.98	.86-.98
NC_{ctt} and $\hat{N}C_{mirt}$	Mean	.9680	.9791	.9821
	Median	.97	.98	.98
	Mode	(.97)- 36	(.98)-49	(.98)-49
	Stand. Dev.	.0217	.0096	.0068
	Range	.86-.99	.93-.99	.97-.99
NC_{ctt} and $\hat{N}C_{mix}$	Mean	.8987	.9114	.9192
	Median	.93	.92	.92
	Mode	(.95)- 19	(.93)-18	(.92)-17
	Stand. Dev.	.0777	.0412	.0301
	Range	.58-.98	.78-.97	.84-.97
NC_{wctt} and $\hat{N}C_{ctt}$	Mean	.8980	.9202	.9161
	Median	.93	.92	.92
	Mode	(.95,.94)-15	(.91)-16	(.90,.91)-15
	Stand. Dev.	.0710	.0337	.0258
	Range	.60-.97	.84-.97	.86-.96
NC_{wctt} and $\hat{N}C_{wctt}$	Mean	.8984	.9195	.9313
	Median	.93	.93	.94
	Mode	(.95)-21	(.93)-23	(.94)-27
	Stand. Dev.	.0747	.0344	.0245
	Range	.60-.97	.80-.97	.86-.97

Table 3 continued

NC_{wctt} and $\hat{N}C_{mirt}$	Mean	.9556	.9637	.9670
	Median	.97	.97	.97
	Mode	(.97)- 40	(.97)-49	(.97)-54
	Stand. Dev.	.0530	.0118	.0088
	Range	.64-.99	.92-.98	.94-.98
NC_{wctt} and $\hat{N}C_{mix}$	Mean	.8945	.8989	.9109
	Median	.92	.91	.91
	Mode	(.96)- 16	(.91)-16	(.90,.91)-16
	Stand. Dev.	.0793	.0479	.0304
	Range	.57-.97	.76-.97	.84-.97
NC_{mirt} and $\hat{N}C_{ctt}$	Mean	.8632	.9093	.9235
	Median	.88	.92	.93
	Mode	(.92)- 9	(.95)-14	(.96)-17
	Stand. Dev.	.0833	.0431	.0324
	Range	.58-.96	.78-.97	.81-.98
NC_{mirt} and $\hat{N}C_{wctt}$	Mean	.8732	.9136	.9293
	Median	.90	.93	.94
	Mode	(.94)- 11	(.95)-14	(.95)-17
	Stand. Dev.	.0858	.0473	.0364
	Range	.59-.96	.78-.98	.78-.98
NC_{mirt} and $\hat{N}C_{mirt}$	Mean	.9335	.9638	.9694
	Median	.95	.97	.97
	Mode	(.96)- 25	(.97)-41	(.97)-44
	Stand. Dev.	.0623	.0117	.0087
	Range	.60-.98	.92-.98	.94-.98
NC_{mirt} and $\hat{N}C_{mix}$	Mean	.8533	.8939	.9114
	Median	.87	.905	.91
	Mode	(.93)- 8	(.94)-13	(.91)-14
	Stand. Dev.	.0912	.0539	.0385
	Range	.55-.96	.72-.98	.79-.98
NC_{mix} and $\hat{N}C_{ctt}$	Mean	.8817	.9144	.9310
	Median	.91	.92	.94
	Mode	(.94)-13	(.94)-15	(.94)-22
	Stand. Dev.	.0812	.0325	.0259
	Range	.60-.98	.81-.96	.85-.97

Table 3 continued

NC_{mix} and $\hat{N}C_{wcn}$	Mean	.8930	.9089	.9228
	Median	.925	.92	.93
	Mode	(.94,.95)-15	(.93)-17	(.94)-24
	Stand. Dev.	.0760	.0356	.0239
	Range	.64-.98	.80-.96	.84-.96
NC_{mix} and $\hat{N}C_{mir}$	Mean	.9496	.9672	.9715
	Median	.97	.97	.97
	Mode	(.98)-29	(.97)-41	(.97)-47
	Stand. Dev.	.0629	.0111	.0079
	Range	.65-.99	.93-.98	.95-.99
NC_{mix} and $\hat{N}C_{mix}$	Mean	.8820	.9046	.9140
	Median	.92	.91	.92
	Mode	(.95)-12	(.93)-17	(.91)-17
	Stand. Dev.	.0857	.0414	.0312
	Range	.60-.98	.78-.97	.82-.97

Figures 1, 3, 5, 7, 9 and 11 represent the distribution of correlations for relationships between true number correct scores. Figures 2, 4, 6, 8, 10, and 12 represent the corresponding distribution of correlations for the relationships between the estimated number correct scores. Each Figure (1-12) represents the distribution of the correlations for all three test lengths: 15-item, 30-item and 50-item. The line with the rhombus symbol represents the 15- item test. The line with the square symbol represents the 30-item test and the line with the triangle symbol represents the 50-item test. As expected, the figures that represent the relationships between estimated number correct scores have more variability than the figures that represent the relationships between the true number correct scores (see Table 1 and Table 2).

1

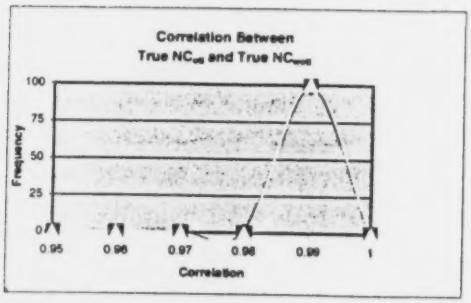

Figure 1.
Correlation of True NC_{cn} and True NC_{wctt}

2

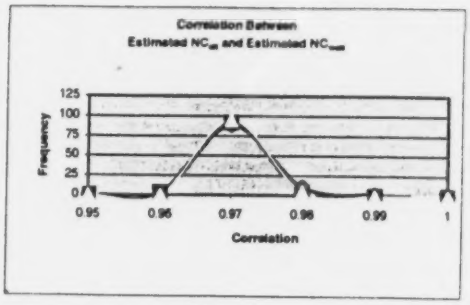

Figure 2.
Correlation of Est. NC_{cn} and Est. NC_{wctt}

3

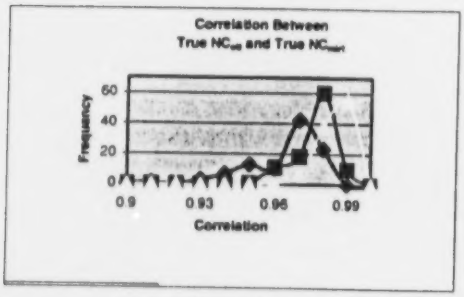

Figure 3.
Correlation of True NC_{cn} and True NC_{mctt}

4

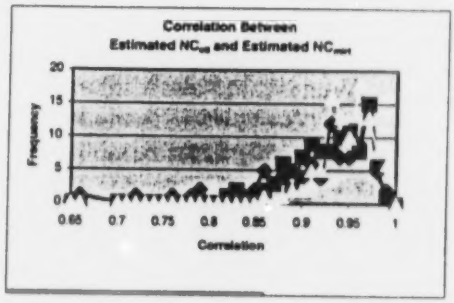

Figure 4.
Correlation of Est. NC_{cn} and Est. NC_{mctt}

5

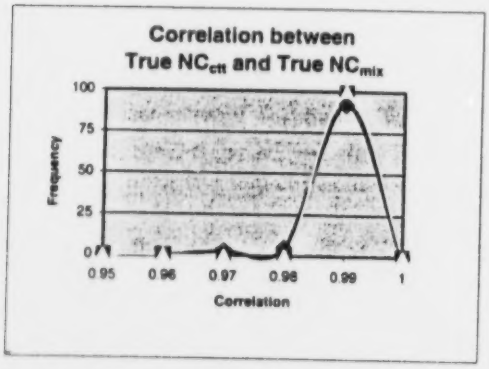

Figure 5.
Correlation of True NC_{cn} and True NC_{mix}

6

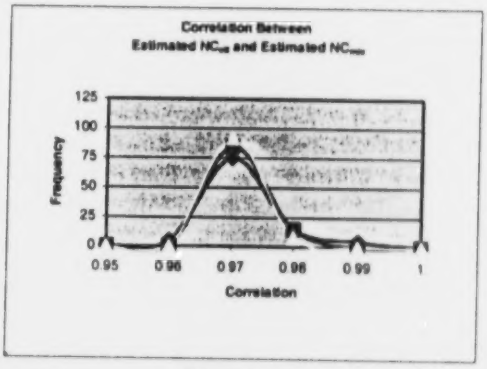

Figure 6.
Correlation of Est. NC_{cn} and Est. NC_{mix}

BEST COPY AVAILABLE

7

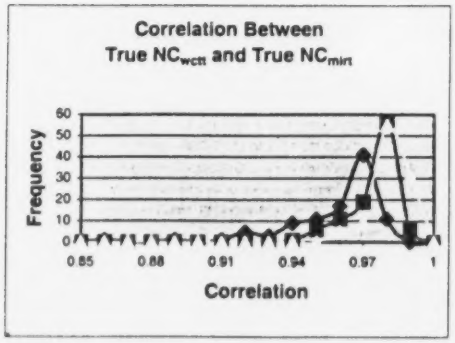

Figure 7.
Correlation of True NC_{wcn} and True NC_{mirt}

8

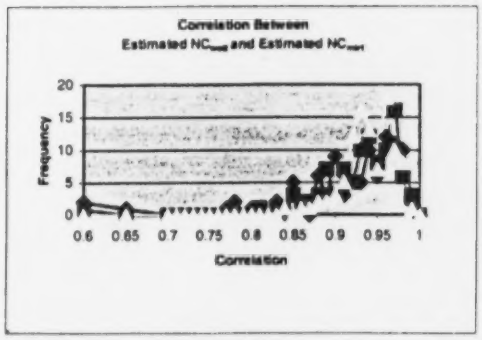

Figure 8.
Correlation of Est. NC_{wcn} and Est. NC_{mirt}

9

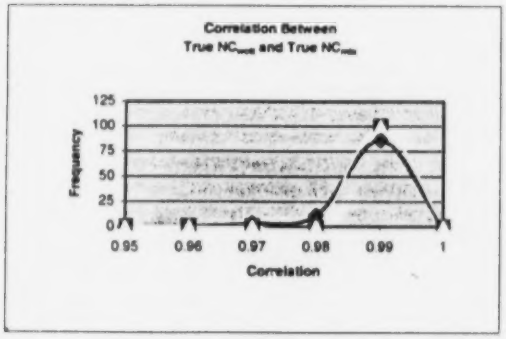

Figure 9.
Correlation of True NC_{wcn} and True NC_{mix}

10

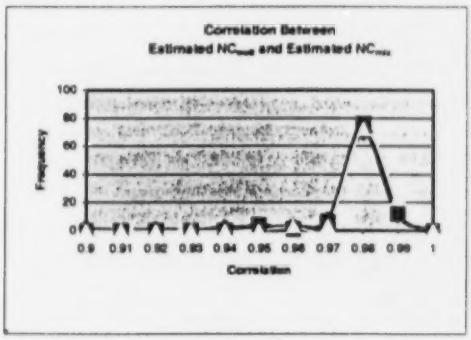

Figure 10.
Correlation of Est. NC_{wcn} and Est. NC_{mix}

11

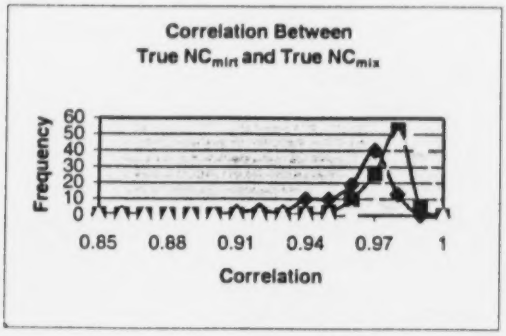

Figure 11.
Correlation of True NC_{mirt} and True NC_{mix}

12

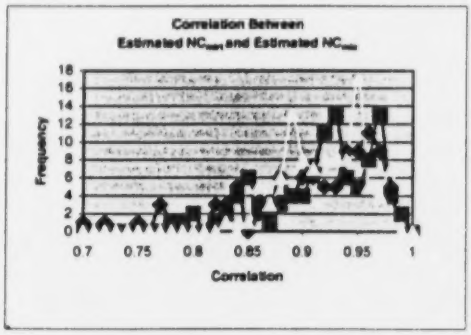

Figure 12.
Correlation of Est. NC_{mirt} and Est. NC_{mix}

Coefficient of Variation (CV)

Table 4 presents the results of the coefficient of variation for the 100 repetitions for each test length. In particular, it shows the number of samples for each method that had the smallest coefficient of variation. The MIX method provides 151 samples that have the smallest coefficient of variation for all the test lengths as compared to 149 samples for the other three methods combined. Thus, the MIX method provides more samples that on average have their estimated scores closer to their true scores than any of the other three methods (CTT, WCTT and MIRT individually or combined). Particularly, for the 15 item test the MIX method had 51 samples that had the smallest coefficient of variation, the MIRT method ranks as the second best method (23 samples), the WCTT method ranks as third (18 samples) and finally the CTT has the least number of samples that have the smallest coefficient of variation (11 samples out of 100). For the 30-item test, the MIX method has almost half of the samples that have the smallest coefficient of variation (49 samples out of 100 samples). The second best method for the 30-item test was the MIRT with 20 samples, the third best method was the CTT with 17 samples and finally the WCTT was the fourth method with only 14 samples. For the 50-item test, the MIX method had more than half of the samples that had the smallest coefficient of variation (54 samples) and the WCTT method ranks as the second best method that had 21 samples. The CTT (13 samples) and MIRT (12 samples) ranked as the third and fourth methods, respectively.

Table 4
Coefficient of Variation for Scoring Methods and Test Lengths

	Test Length		
Method	15 Items	30 Items	50 Items
CTT	11 (4)	17(3)	13(3)
WCTT	18 (3)	14(4)	21(2)
MIRT	23 (2)	20(2)	12(4)
MIX	51 (1)	49(1)	54(1)
TOTAL	100	100	100

The indices in table 1 represent the number of samples that had the smallest Coefficient of Variation (CV). The numbers in the parenthesis represent the rank of each scoring method for each test length. The method with the most samples that have the smallest CV ranks as one and the method with the least number of samples that have the smallest CV ranks as fourth.

In summary, the MIX method has the most samples (about 50 % of the samples) that have the smallest coefficient of variation across the three test lengths (15-item, 30-item test and 50-item test). Based on Table 4, the remaining three methods (CTT, WCTT and MIRT) rank in a different order for each of the test lengths.

C. Bootstrap Analysis.

Bootstrap techniques were employed to test for significant differences between the means of the CV for the four number correct scoring methods on the 50-item test. Ninety-five percent confidence intervals of the mean of the coefficient of variations for the 50-item test for the four scoring methods were formed from 1000 bootstrap repetitions (B=1000). The results of the Bootstrap Analysis show that the MIX method was statistically significantly different from the CTT method, WCTT method and the MIRT method. Also, the CTT method was statistically significantly different from the MIRT method. Thus, the number correct based on the MIX method has the smallest mean of the CVs. Table 5 presents the results of the Bootstrap Analysis.

Table 5

Bootstrap confidence intervals of the coefficient of variation.
The intervals for the 4 methods are based on 1000 replications of the 50 item test.

Scoring Method	95% Confidence Interval	Confidence Bands .7	.8	.9
MIX	[.70, .75]	xxxxxxx		
CTT	[.76, .81]		xxxxxxx	
WCTT	[.76, .82]		xxxxxxxx	
MIRT	[.82, .87]			xxxxxxxx

5. Summary

The *MIX* method of Number Correct Scoring was the most accurate of the four method used to estimate the true score of an examinee. The *MIX* method had many more samples that had the smallest coefficient of variation than any of the other three number correct scoring methods (considered separately). This outcome was consistent for all three-test lengths (15-item, 30-item and 50-item). Finally, the *MIX* method was significantly different from the other three scoring methods, *CTT, WCTT,* and *MIRT* using the bootstrap analysis .

Appendix

Notation

1. θ_1 = Ability parameter of an examinee in dimension one will be generated using an algorithm coded in Fortran 77. Subroutines were Uni1 and Normb1(Blair, 1987).
2. θ_2 = Ability parameter of an examinee in dimension two will be generated using an algorithm coded in Fortran 77. Subroutines were Uni1 and Normb1 (Blair, 1987).
3. $\hat{\theta}_1$ = Estimated ability of an examinee in dimension one will be calculated using the TESTFACT program (Muraki and Engelhard, 1985).
4. $\hat{\theta}_2$ = Estimated ability of an examinee in dimension two will be calculated using the TESTFACT program (Muraki and Engelhard, 1985).
5. a_1 = Item discrimination parameter in dimension one will be generated using an algorithm coded in Fortran 77.
6. a_2 = Item discrimination parameter in dimension two will be calculated using equation (2).
7. $\hat{a}_1$ = Estimated item discrimination parameter for dimension one will be calculated using the TESTFACT program.
8. $\hat{a}_2$ = Estimated item discrimination parameter for dimension two will be calculated using the TESTFACT program.
9. d_i = Item discrimination defined under CTT. This value will be computed using the item/total-test score point biserial correlation.

10. $\hat{d}_i$ = Estimated item discrimination value of an item under CTT. This value will be calculated using the TESTFACT program.

11. NC_{ctt} = True number correct score of an examinee, based on the traditional method under the Classical True-score Theory (CTT).

12. $\hat{N}C_{ctt}$ = Estimated number correct score for an examinee using the traditional method under CCT is the sum of the ones for a given row of the $N\hat{x}K_{ctt}$ matrix.

13. NC_{wctt} = True number correct score of an examinee, based on items that are weighted according to their discrimination value as defined under CTT.

14. $\hat{N}C_{wctt}$ = Estimated number correct score for the items that are weighted according to their discrimination value defined in CTT.

15. NC_{mirt} = True number correct score under Multidimensional Item Response Theory (MIRT)

16. $\hat{N}C_{mirt}$ = Estimated number correct score under Multidimensional Item Response Theory.

17. NC_{mix} = True number correct score of an examinee, based on items that are weighted according to the item parameters defined in MIRT and test scores based on CTT. This is called the MIX method.

18. $\hat{N}C_{mix}$ = Estimated number correct score using the MIX method.

References

Ackerman. T. A. (1994). Graphical representation of multidimensional item response theory analyses. New Orleans: Paper presented at the annual meeting of the American Educational Research Association.

Allen, M. and Yen, W. (1979). Introduction to Measurement Theory. Brooks/Cole Publishing Co. Monterey, CA.

Ansley, T. N. and Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. Applied Psychological Measurement. Vol 9, No. 1, pp 37-48.

Blair, R. C. (1987) Rangen. Boca Raton, Fl, IBM.

Birnbaum, A. (1968). Some latent models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (1968). Statistical Theories of Mental Test Scores. (Chapters 17-20), Reading, MA: Addison-Wesley.

Efron, B, & Tibshirani, R. J. (1998) An introduction of the bootstrap. Boca Raton, FL: CRC.

Frary, B. R. (1989). Partial-Credit Scoring Methods for Multiple-Choice Tests. Applied Measurement in Education. Vol 2, No 1, pp. 79-96.

Hampleton, R., Swaminathan, H. and Rogers, J. (1991). Fundamentals of item response theory. London: SAGE.

Howell, D. (1997). Statistical Methods for Psychology. (Fourth Edition). Belmont, CA: Wadsworth.

Lahey, Computer Systems, Inc. (1994). Personal Fortran (version 3.0). Incline Village, NV: Author.

Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Erlbaum.

Muraki, E. and Englelhard, G. (1985). Full-Information Item Factor-Analysis: Application of EAP Scores. Applied Psychological Measurement. Vol. 9, No. 4, pp. 417-430.

Pierce, D. (June, 1985). F-M High School Counseling Center. Standardized Tests/Preparation. [Online]. Available: http://www.fmhs.cnyric.org/Fmcoun/Fmcoun/standard.htm

Reckase, M. D. Development and application of a multivariate logistic latent trait model. (Doctoral dissertation, Syracuse University, 1972). Dissertation Abstracts International, 1973, 33. (University Microfilms No. 73-7762)

Reckase, M. D. (1986). The discriminating power of items that measure more than one dimension. San Francisco: Paper presented at the annual meeting of the American Educational Research Association.

Reckase, M. D. and McKinley, R. L. (1986). Some latent theory in a multidimensional latent space. (ERIC Document Reproduction Service No. ED264265)

Stocking, L. M. (1996). An alternative method for scoring adaptive tests. Journal of Educational and Behavioral Statistics. Vol 21, No 4, pp 365-389.

Way, W. D., Ansley, T. N. and Forsyth, R. A. (1988). The comparative effects of compensatory and noncompesatory two-dimensional data on unidimensional IRT estimates. Applied Psychological Measurement. Vol 12, No. 3, pp 239-252.

Weaver, C. (1995). Facts. On standardized tests and assessment alternatives. [Online]. Available: http://www.heinemann.com/info/08894f10.html

Wilson, T. D., Wood, R. and Gibbons R. (1998). TESTFACT. Scientific Software International Incorporated.